Structure learning with independent non-identically distributed data

Robert E. Tillman RTILLMAN@CMU.EDU
Department of Philosophy and Machine Learning Department, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA

Abstract

There are well known algorithms for learning the structure of directed and undirected graphical models from data, but nearly all assume that the data consists of a single i.i.d. sample. In contexts such as fMRI analysis, data may consist of an ensemble of independent samples from a common data generating mechanism which may not have identical distributions. Pooling such data can result in a number of well known statistical problems so each sample must be analyzed individually, which offers no increase in power due to the presence of multiple samples. We show how existing constraint based methods can be modified to learn structure from the aggregate of such data in a statistically sound manner. The prescribed method is simple to implement and based on existing statistical methods employed in metaanalysis and other areas, but works surprisingly well in this context where there are increased concerns due to issues such as retesting. We report results for directed models, but the method given is just as applicable to undirected models.

1. Introduction

Directed and undirected graphical models are now ubiquitous in machine learning. They are used for a variety of applications including probabilistic and causal inference, prediction, feature selection, data visualization, and modeling interventions. Learning the structure of these models from data has remained an active area of research for several decades. Most existing structure learning algorithms are either constraint based, which use conditional independence tests, or score based, which use decomposable score functions such as BIC. The well known constraint based PC algorithm is guaranteed to learn the correct directed acyclic graphical model (DAG) from data under reasonable assumptions (Spirtes et al., 2000) and is useful for datasets with hundreds of variables in general or thousands with an additional sparcity assumption (Kalisch & Bühlmann, 2007). The GES algorithm is a score-based alternative with the same asymptotic guarantees (Chickering, 2002), but in practice is useful only with datasets that have far fewer variables.

Most existing structure learning algorithms assume that the data consists of a single i.i.d. sample. In some cases, however, we may obtain data from multiple sources which we assume were generated by some common data generating mechanism, but are not necessarily identically distributed. Such data occur frequently in fMRI analysis when a sequence of recordings are made for multiple individuals as they are each subjected to the same experimental conditions. We tend to observe the same dependencies between the measured variables for each individual, but due to differences in individual brains, the distribution of the data can vary significantly across individuals even when the same fMRI machine is used. Similarly, with the increasing availability of large amounts of data, researchers are often able to obtain multiple legacy datasets which measure the variables they are interested in, but may vary in distribution due to differences in recording instruments or slight variations in experimental procedures. Pooling such data can lead to a number of well known statistical problems, such as introducing new spurious associations and Yule-Simpson effects. This can have a significant effect on the accuracy of structure learning algorithms, as we demonstrate in section 3. Standardizing the data to 0 mean and unit variance will not necessarily avoid these errors. In these circumstances, experimenters typically must base their analysis on only one of the samples or analyze each separately and extrapolate a model for the aggregate from these results. This wastes useful data and does not produce the increase in statistical power we typically observe when we obtain more data.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

In this paper, we show how existing constraint based structure learning methods can be modified to learn from the aggregate of such data where samples are obtained from independent sources but may not be identically distributed. The method given is simple to implement and based on existing statistical methods employed in metaanalysis and other other areas to address simpler problems. We show that despite the increased fragility of these methods when they are used to learn structure, due to retesting the same data and results of early independence tests influencing which tests are performed later, significant increases in accuracy are observed for these methods when compared to using a single sample or the standardized pooled data. In some instances, accuracy is greater than when an i.i.d. sample of the same size as the aggregate data is used. Section 2 describes this method and the relevant background. Section 3 then presents experimental results which compare the method to using a single i.i.d. sample and the standardized pooled data. Section 4 briefly discusses the related problem of learning structure from distributed data with overlapping variables and shows how this method is useful for this problem. Conclusions and possible future research are discussed in section 5.

2. Learning with combined p-values

2.1. Background

Metaanalyses are statistical summaries of (possibly contrary) experimental findings. Statistical procedures are used to combine results from multiple independent experiments and obtain a global assessment of whether certain effects may be present in the population. Experimenters typically report p-values associated with effects that are observed (or not observed), which indicate probabilities of attaining a test statistic as or more extreme than the observed statistic under the null hypothesis no effect is present. Thus, a common and convenient method for fusing experimental results is to combine the observed p-values from each experiment into an appropriate new statistic which can be used to decide whether the null hypothesis should be rejected. If we know the distribution of the new statistic, we can obtain a new p-value using the observed value, which is representative of the data from all of the studies. We refer to such p-values as *combined p-values*. Such methods are used regularly in metanalysis (Sutton et al., 2000) and also for certain tasks in neuroimaging (Lazar et al., 2002). A significant advantage of these methods over some other data aggregation procedures is that they are not affected by differences in distribution across the different experiments, since they rely only on reported p-values.

The most naive method for combining information from multiple p-values is to simply average the individual p-values. This, however, ignores the fact that each p-value constitutes an independent source of information and unlikely evidence from multiple sources is more compelling than unlikely evidence from a single source. For example, assume we set α, our acceptable type I error rate, to .05 and observe $p_1 = p_2 = .06$ in two independent studies. The average p-value is .06 even though it is much less likely that we would observe p-values this extreme in two independent studies than in a single study. We should expect a combined p-value below the α threshold, which would change the decision we make about rejecting the null hypothesis.

More sophisticated methods have thus been proposed for combining information from multiple p-values that account for the independent sources of evidence in a theoretically sound manner. One of the most well known is due to R.A. Fisher (Fisher, 1950). Fisher's method requires computing the statistic T_F which, as shown below, has a χ^2 distribution with $2k$ degrees of freedom under the null hypothesis, where k is the number of p-values combined.

$$T_F = -2\sum_{i=1}^{k}\log(p_i) \tag{1}$$

$$\stackrel{d}{=} \sum_{i=1}^{k}\frac{1}{2}\exp\left(\frac{-(-2\log(p_i))}{2}\right) \tag{2}$$

$$\stackrel{d}{=} \frac{1}{\Gamma(k)2^k}T_F^{k-1}\exp\left(\frac{-T_F}{2}\right) \tag{3}$$

$$\stackrel{d}{=} \frac{1}{\Gamma\left(\frac{2k}{2}\right)2^{\frac{2k}{2}}}T_F^{\frac{2k}{2}-1}\exp\left(\frac{-T_F}{2}\right) \tag{4}$$

$$\sim \chi^2_{2k} \tag{5}$$

This follows simply from the fact that p-values have a standard Uniform distribution under the null hypothesis, which becomes an Exponential distribution after the log transformation. The sum of k Exponentially distributed random variables has a Gamma distribution, of which the χ^2 is a special case. After calculating T_F, we can obtain a combined p-value for k studies, which we denote $\tilde{p}$, by calculating $1 - F_{\chi^2_{2k}}(T_F)$, where $F_{\chi^2_{2k}}$ is the χ^2_{2k} cumulative distribution function (or compare the observed value of T_F to the $1-\alpha$ χ^2_{2k} quantile to make the rejection decision). Figure 1 shows the resulting $\tilde{p}$ when Fisher's method is used to combine two p-values, as the two p-values are varied from 0 to .25. It is easy to see that obtaining two low p-values results in a much lower $\tilde{p}$, unlike taking the average. In simulations and many previous studies,

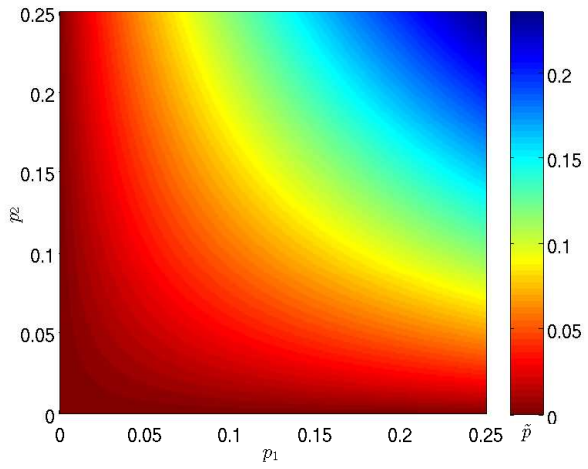


Figure 1. Fisher combined p-value for two p-values

Fisher's method has in general been more reliable than other similar methods for combining p-values (Lazar et al., 2002). In addition, Fisher's method satisfies an optimality criterion known as *Bahadur efficiency* (Bahadur, 1971), which is related to the effective use of data as the number of samples increases (Lazar et al., 2002).

We now briefly describe some of the most common competing methods. Tippett (1950) and Worsely and Friston (2000) propose using the first and kth order statistics, respectively, for $\tilde{p}$,

$$T_T = \min_{i=1}^{k} p_i \qquad T_{WF} = \max_{i=1}^{k} p_i$$

but require that the original α be transformed to $1-(1-\alpha)^{1/k}$ and $\alpha^{1/k}$, respectively. Stouffer et al. (1949) uses the following test statistic, where Φ^{-1} is the inverse Gaussian cumulative distribution function.

$$T_S = \sum_{i=1}^{k} \frac{\Phi^{-1}(1-p_i)}{\sqrt{k}}$$

Under the null hypothesis, T_S has a standard Gaussian distribution. For a two-tailed test, we compute the combined p-value by finding $2P(X \geq |T_S|)$, and use the original α level. Mudholkar and George (1979) uses a logit transform of the p-values to obtain the following statistic, where $c = \sqrt{3(5k+4)/k\pi^2(5k+2)}$.

$$T_{MG} = -c \sum_{i=1}^{k} \log\left(\frac{p_i}{1-p_i}\right)$$

Under the null hypothesis, T_{MG} has a standard t distribution with $5k+4$ degrees of freedom. For a two-tailed test, we compute the combined p-value by finding $2P(X \geq |T_{MG}|)$ and use the original α level. Like Fisher's method, Mudholkar and George's method is Bahadur efficient (Lazar et al., 2002).

2.2. Application to structure learning

Constraint based structure learning algorithms interact with the data only when doing conditional independence tests. Since conditional independence tests are simply special types of hypothesis tests, they can be replaced with combined p-value tests when we have multiple (k) samples which we believe are from same data generating mechanism, but which may have different distributions.[1] We propose the following method for modifying constraint based algorithms to take advantage of such data: whenever the constraint based algorithm queries whether $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, we (1) find p-values $p_1, \ldots, p_k$ under the null hypothesis $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ using the data from each of the k samples and an appropriate independence test; (2) obtain $\tilde{p}$ for $p_i, \ldots, p_k$ using one of the methods described in section 2.1; and (3) reject $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ if $\tilde{p} < \alpha$, our chosen allowable type I error rate, and report this result to the constraint based algorithm.

No additional theoretical justifications are required for this modification. As more data are acquired (more samples or more data points in each sample) it becomes increasingly unlikely that the test statistic will land in the rejection region under the null hypothesis, just as in the standard case. Practical performance with this modification, however, is not at all obvious. Many statistical issues come up in structure learning that do not arise in the simpler cases in which these methods have previously been evaluated. Many independence tests over the same variables are required to learn structure so data is often retested. Furthermore, the PC algorithm and many other structure learning algorithms use greedy procedures to limit the number of tests performed. The results of early tests decide which additional tests needs to be performed. Errors made in early tests can have a more significant effect on accuracy. In the next section, we thus empirically investigate the practical performance of this modification for each of the methods described in section 2.1.

3. Experimental results

The PC algorithm is one of the most well known constraint based DAG structure learning algorithm and

[1] It is important to emphasize that we are *not* referring to samples from different treatment conditions in an experiment. The differences in distributions should only be due to factors such as different response rates in individuals, differences in measurement instruments, and minor differences in the experimental procedures.

has been used as a benchmark to compare to many other structure learning algorithms. In the experiments below, we evaluate modified versions of PC which use one of the combined p-value tests described in section 2.1. We compare these structure learning procedures for multiple samples to the baseline performance of the standard PC algorithm using (a) only one of the samples and (b) a concatenated dataset formed by pooling together each sample after standardizing to mean 0 and unit variance. We follow the usual distribution conventions found in the structure learning literature when generating synthetic data: we use unconstrained multinomial distributions for discrete variables and linear Gaussian distributions for continuous variables (Chickering, 2002). We use Pearson's χ^2 conditional independence test for discrete data and Fisher's Z test for continuous data, and set $\alpha = .05$.

We first consider the case of learning a DAG structure from two non-identically distributed samples. We generated 100 random DAG structures with 15 nodes, and 100 random DAG structures with 40 nodes, using the Melançon et al. (2000) MCMC algorithm, to represent a range of possible DAG structures we might learn from data. Random parameters for both discrete and continuous distributions were generated twice for each structure, i.e. we specify two non-identical discrete distributions and two non-identical continuous distributions for each structure. Forward sampling was used to generate datasets of sizes (N) 50, 100, 500, 1000, and 2500 for each structure and distribution. For each structure and sample size, we used the modified and baseline PC procedures described above to learn structure from the corresponding pairs of discrete datasets with different distributions and continuous datasets with different distributions. Figures 2, 3, 4, and 5 report the average (over the 100 random structures) number of edge omission, edge commission, and orientation (reversed edges) errors for the 15 variable discrete, 15 variable continuous, 40 variable discrete, and 40 variable continuous cases, respectively. Error bars indicate 95% confidence intervals. For edge omissions, a noticeable increase in accuracy is observed in nearly every evaluation with each of the combined p-value methods, except for Worsely and Friston's method, when compared to using a single dataset or the concatenated data. Fisher's method outperforms the other methods for combining p-values, which is consistent with the performance of these methods in other applications (Lazar et al., 2002). In some cases, Fisher's method, when applied to two datasets of the same size actually outperforms the single dataset method when applied to a dataset of twice that size, e.g. N=50 vs. N=100. This indicates that it can be more advantageous to use this method with two independent smaller samples than with one larger sample. All of the methods except the concatenated data make few edge commissions; the large number of edge commissions with the concatenated data indicates that many spurious associations result when the non-identically distributed data is pooled. The combined p-value methods also show fewer orientation errors for the 15 variable continuous cases with sample sizes 500, 1000, and 2500. Fisher's method again shows the fewest errors. The runtimes for each method are shown in figure 6. A small increase in runtime is observed for the combined p-value methods when compared to using a single dataset. These runtimes, however, remain within one second in the worst case for continuous data and five seconds for discrete data. PC sometimes becomes very slow with the concatenated data due to increased statistical errors.

We also tested these methods with well known benchmark Bayesian networks (DAG structures with associated joint probability distributions over the DAG nodes). The Alarm network contains 37 nodes, 46 edges, and has maximum node degree (edges with endpoints at a particular node) 6. The Insurance network contains 27 nodes, 52 edges, and has maximum node degree 9. The Hailfinder network contains 56 nodes, 66 edges, and has maximum node degree 17. Each network has a specified discrete parametrization. For each network, we generated 100 samples of size 2500 from the given parametrization and 100 samples of size 2500 from a random discrete parametrization. Table 1 shows average edge omission, edge commission, and orientation errors and runtimes with 95% confidence intervals for each of the methods from the previous experiment. Fisher's method again outperforms the other methods for combining p-values and shows a significant improvement when compared to using a single dataset or the concatenated data, especially for the Hailfinder network where many of the other methods for combining p-values perform poorly. We again observe a slight increase in runtime when combining p-values, but this increase is not particularly large relative to the size and difficulty of learning the networks.

Finally, to confirm that the observed trend continues as we increase the number of samples with different distributions, we held the sample size fixed at 100 and repeated the the first experiment as we varied the number of samples with different distributions. The results for the discrete and continuous cases were nearly identical so we report errors only for the 15 discrete variables case in figure 7.

As expected, these results indicate that as we increase

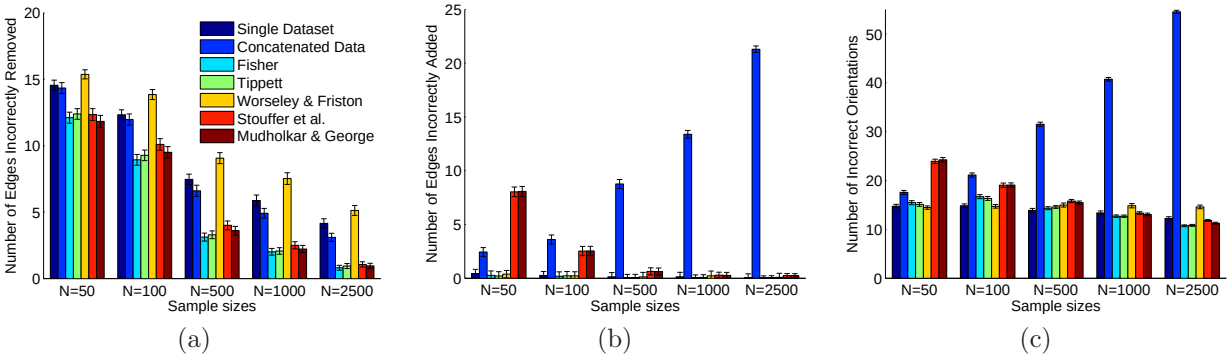


Figure 2. (a) edge omissions, (b) edge commissions, and (c) orientation errors with 15 discrete variables

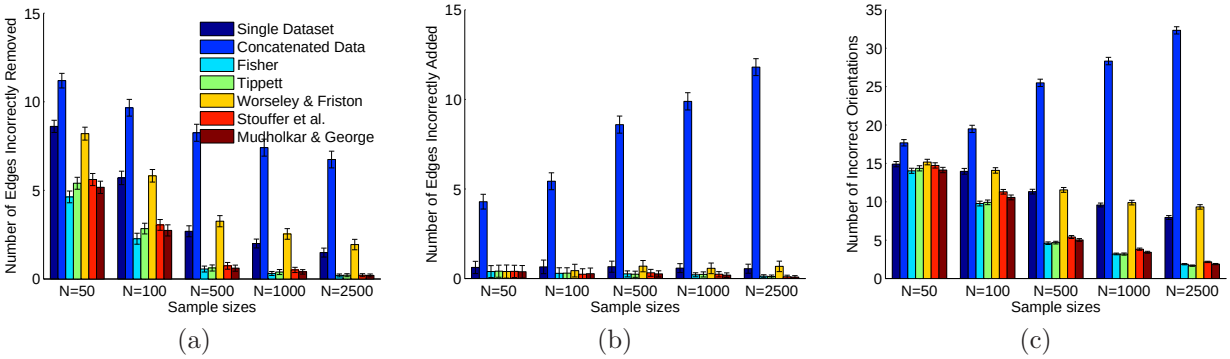


Figure 3. (a) edge omissions, (b) edge commissions, and (c) orientation errors with 15 continuous variables

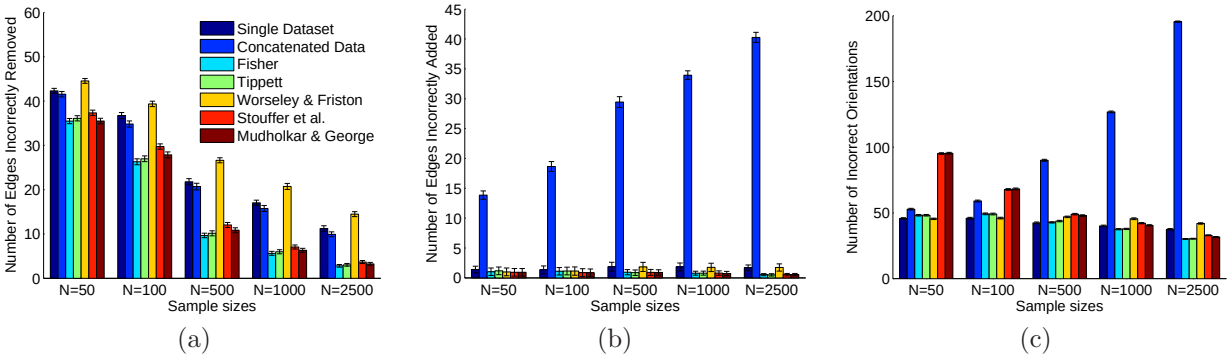


Figure 4. (a) edge omissions, (b) edge commissions, and (c) orientation errors with 40 discrete variables

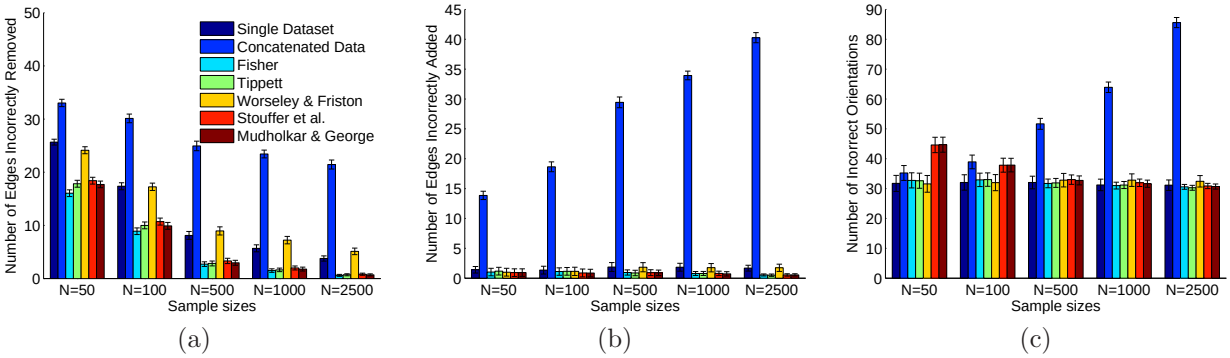
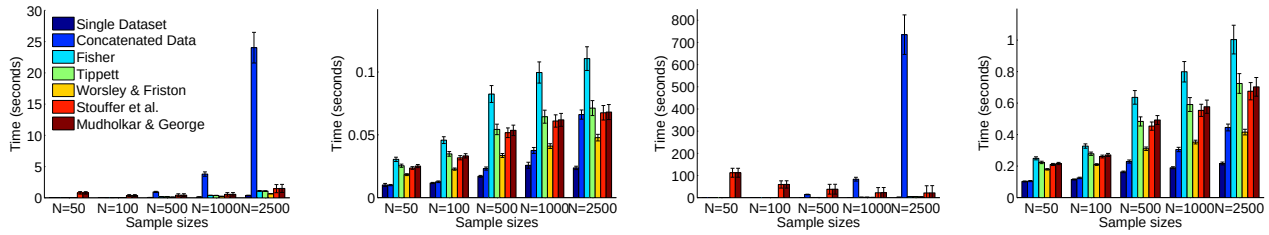


Figure 5. (a) edge omissions, (b) edge commissions, and (c) orientation errors with 40 continuous variables



(a) (b) (c) (d)

Figure 6. Runtimes with (a) 15 discrete variables (b) 15 continuous variables, (c) 40 discrete variables, and (d) 40 continuous variables

Table 1. Errors for the (a) Alarm, (b) Insurance, and (c) Hailfinder networks

	Method	Edge Omissions	Edge Commissions	Orientations	Time (seconds)
	Single Dataset	5.00 ± 0.00	2.00 ± 0.00	17.00 ± 0.00	3.01 ± 0.02
	Concatenated Data	1.72 ± 0.27	123.68 ± 6.42	257.55 ± 12.27	1419.92 ± 114.55
	Fisher	0.73 ± 0.16	1.48 ± 0.15	12.50 ± 1.01	6.92 ± 0.05
(a)	Tippett	0.73 ± 0.15	1.87 ± 0.15	13.73 ± 1.02	7.05 ± 0.05
	Worsely & Friston	6.18 ± 0.37	0.15 ± 0.11	25.67 ± 1.03	4.37 ± 0.70
	Stouffer et al.	1.30 ± 0.19	1.95 ± 1.10	12.85 ± 1.26	65.26 ± 44.65
	Mudholkar & George	0.95 ± 0.16	2.30 ± 1.08	13.38 ± 1.35	65.44 ± 44.62
	Single Dataset	16.00 ± 0.00	1.00 ± 0.00	39.00 ± 0.00	3.27 ± 0.09
	Concatenated Data	1.69 ± 0.26	109.78 ± 2.82	257.20 ± 5.18	1101.25 ± 194.36
	Fisher	2.22 ± 0.21	0.20 ± 0.09	38.52 ± 0.82	19.90 ± 1.93
(b)	Tippett	2.44 ± 0.25	0.20 ± 0.09	38.90 ± 0.89	20.26 ± 1.92
	Worsely & Friston	15.93 ± 0.22	0.56 ± 0.13	41.34 ± 0.79	$5.26 \pm .48$
	Stouffer et al.	3.17 ± 0.21	0.33 ± 0.11	42.87 ± 0.83	17.91 ± 1.62
	Mudholkar & George	2.62 ± 0.25	0.35 ± 0.12	41.72 ± 0.84	18.72 ± 1.76
	Single Dataset	34.00 ± 0.00	33.00 ± 0.00	76.00 ± 0.00	14.23 ± 0.33
	Concatenated Data	3.97 ± 0.28	113.60 ± 2.36	235.40 ± 5.14	1509.16 ± 191.80
	Fisher	4.15 ± 0.22	2.53 ± 0.22	25.45 ± 1.86	313.65 ± 23.55
(c)	Tippett	4.15 ± 0.20	2.83 ± 0.13	27.50 ± 1.68	313.66 ± 25.22
	Worsely & Friston	33.90 ± 0.14	36.73 ± 0.72	83.73 ± 1.18	53.46 ± 11.10
	Stouffer et al.	4.08 ± 0.25	60.70 ± 0.46	106.55 ± 1.02	$928.66.4 \pm 41.77$
	Mudholkar & George	3.88 ± 0.23	62.37 ± 0.50	107.78 ± 0.90	$924.56.3 \pm 50.91$

the number of samples, fewer errors are made by each of the combined p-value methods except for Worsely and Friston's, and the computational tradeoff is small.

These results show that the procedure described in section 2.2 can lead to noticeable gains in accuracy when learning structure from data with different distributions when compared to using only one sample or pooling the standardized data. Fisher's method outperforms the others in all of the evaluations so we recommend it in general. If a researcher has prior knowledge about the data, then there may be a reason to use another method. Such considerations, however, are beyond the scope of this paper. A comprehensive investigation of each method may yield information about specific contexts where each may be optimal.

We also note the overall poor performance of Worsely and Friston's method. Both Worsely and Friston's and Tippett's methods throw away all p-values except the max and min, respectively. Unlike the other methods, these are not based on sufficient statistics so we should expect poor performance. The better performance of Tippett's method is most likely artificial.

4. Relation to learning from distributed data with overlapping variables

Tillman et al. (2009) and Tillman (2008) investigate learning structure from distributed data with overlapping variables, i.e. collections of datasets with some variables in common, but other variables unique to

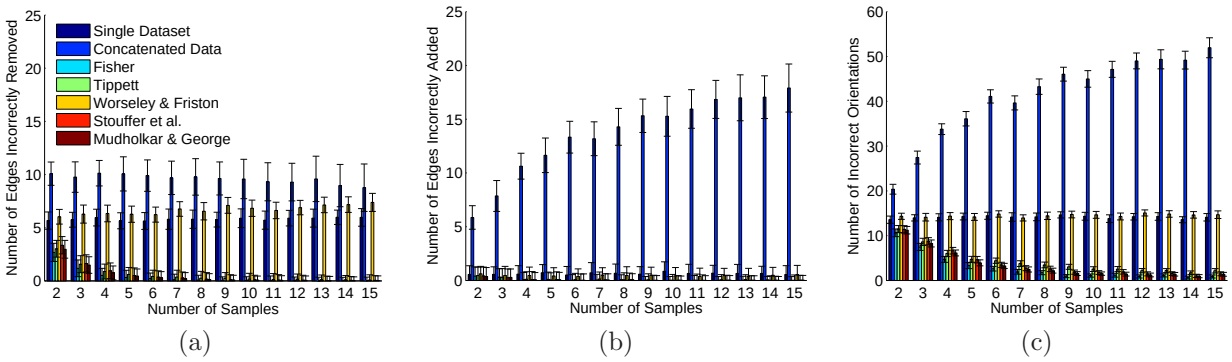


Figure 7. (a) edge omissions, (b) edge commissions, and (c) orientation errors for 15 discrete variables

particular datasets. An example of this is econometric data for the United States and United Kingdom economies. Due to differences in financial recording conventions, the U.S. and U.K. record different sets of variables, but many of the recorded variables are recorded by both countries, i.e. they are overlapping variables. Constraint based algorithms for this scenario have been proposed which learn the complete set of structures that are consistent with every dataset. Since sets of independencies can imply further independencies, e.g. through the graphoid axioms (Pearl, 2000), these algorithms can exclude many more structures than only those that are inconsistent with the conditional independencies common to every dataset.

Due to statistical errors, conditional independence tests for particular variables can yield contrary results for two or more datasets where we should expect the same independencies. This poses a major problem for these algorithms and is often undetectable in practice, e.g. a learned association may conflict with an independence that is implied by two independence decisions made at different stages in the algorithm. Even a considerable amount of bookkeeping cannot avoid this problem since deriving every independence implied by a set of independencies can require an infinite sequence of steps (Studený, 1992). Undiscovered contradictory independence information causes these algorithms to find no structures that are consistent with the data. This problem occurs more frequently as the dimensionality of the data increases since conditional independence tests become less reliable. As a result, the algorithms have thus far only been useful for datasets with a few variables. We found that with 10 variables and $N = 2500$, the DCI algorithm, described in (Tillman, 2008), returns structures consistent with the data less than half of the time (out of 100 examples). Figure 8 plots the trend.

We combined the Fisher combined p-value test with the DCI algorithm and had much better results. In our modified algorithm, whenever DCI would normally perform a conditional independence test using a specific dataset, we checked whether all of the variables involved in the test were contained in another dataset and if so used the Fisher combined p-value tests with each of these datasets to make the independence decision.[2] Consistent structures were returned every time for 100 random 15 variable discrete and continuous cases with each sample size shown in figure 8, which indicates that this method prevents these errors from occurring 100% of the time. We observed edge omission, edge commission, and orientation errors comparable to previous simulations of the DCI algorithm using datasets with fewer variables.

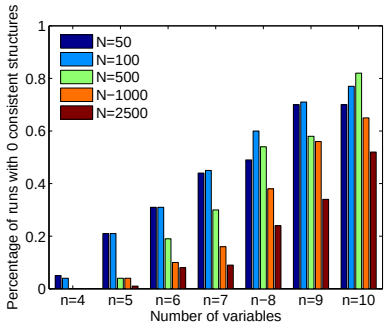


Figure 8. Frequency no consistent structures are found

5. Conclusions

In many contexts where structure learning is relevant, we may encounter data from multiple sources. If we believe that the data are from the same data generating mechanism, but the data from each source are not identically distributed, then we should not naively pool

[2] We actually made further changes to handle more complicated cases, but this is beyond the scope of this paper.

the data to do structure learning, since this can result in significant errors as we demonstrated in section 3. However, relying on a single sample is unsatisfying; it wastes useful data and does not result in the usual increase in statistical power that is observed when we obtain more data. In this paper we described a statistically sound way to learn structure using the aggregate of such data which is easy to implement by modifying existing constraint based procedures. Our experimental results indicate that using this method can produce significant gains in accuracy, especially when Fisher's method is used, without a significant computational tradeoff when compared to using a single sample or the pooled standardized data. These gains in accuracy occur despite concerns of retesting the same data and results of early independence tests determining which later tests are performed, which were not an issue in previous evaluations of the methods given in section 2.1.

As mentioned in section 3, there is room for additional investigations into each of the methods for combining p-values to characterize any instances where each might be favorable to use when learning structure from non-identically distributed data. Considering and comparing possible alternative score based methods for structure learning with non-identically distributed samples is another topic that can be investigated in future research.

Acknowledgments

We thank David Danks, William Eddy, Clark Glymour, Teddy Seidenfeld, and Peter Spirtes for helpful comments and suggestions. Robert E. Tillman was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative.

References

Bahadur, R. R. (1971). *Some limit theorems in statistics*. Philadelphia: SIAM.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507–554.

Fisher, R. A. (1950). *Statistical methods for research workers*. London: Oliver and Boyd. 11th edition.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Lazar, N. A., Luna, B., Sweeney, J. A., & Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, *16*, 538–550.

Melançon, G., Dutour, I., & Bousquet-Mélou, M. (2000). *Random generation of dags for graph drawing* (Technical Report INS-R0005). Centre for Mathematics and Computer Sciences, Amsterdam.

Mudholkar, G. S., & George, E. O. (1979). The logit method for combining probabilities. *Symposium on Optimizing Methods in Statistics* (pp. 345–366). New York: Academic Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, United Kingdom: Cambridge University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press. 2nd edition.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams, R. M. (1949). *The american soldier: Vol. 1. adjustment during army life*. Princeton: Princeton University Press.

Studený, M. (1992). Conditional independence relations have no finite complete characterization. In S. Kubik and J. A. Visek (Eds.), *Information theory, statistical decision functions and random processes*, vol. B, 377–396. Dordrecht: Kluwer.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. New York: John Wiley & Sons.

Tillman, R. E. (2008). *Learning Bayesian network structure from distributed data with overlapping variables* (Technical Report). Carnegie Mellon University, Pittsburgh, PA.

Tillman, R. E., Danks, D., & Glymour, C. (2009). Integrating locally learned causal structures with overlapping variables. *Advances in Neural Information Processing Systems 21*.

Tippett, L. H. C. (1950). *The method of statistics*. London: Williams and Norgate. 1st edition.

Worsley, K. J., & Friston, K. J. (2000). A test for conjunction. *Statistics and Probability Letters*, *47*, 135–140.